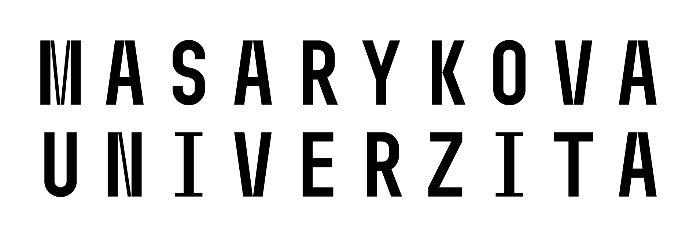


PŘÍRODOVĚDECKÁ FAKULTA

Diplomová práce

**Hoa Vu Thu**

**Brno 2020**



PŘÍRODOVĚDECKÁ FAKULTA

Klasifikace zdravotnických dat prostřednictvím neuronových sítí

Diplomová práce

**Hoa Vu Thu**

Vedoucí práce: RNDr. Martin Komenda, Ph.D.

Název ústavu

**Brno 2020**

**Bibliografický záznam**

|  |  |
| --- | --- |
| **Autor:** | Bc. Hoa Vu Thu  Přírodovědecká fakulta, Masarykova univerzita  Název ústavu |
| **Název práce:** | Klasifikace zdravotnických dat prostřednictvím neuronových sítí |
| **Studijní program:** | Experimentální biologie |
| **Studijní obor:** | Matematická biologie |
| **Vedoucí práce:** | RNDr. Martin Komenda, Ph.D. |
| **Akademický rok:** | 2019/2020 |
| **Počet stran:** | ??+?? |
| **Klíčová slova:** | Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo |

**Bibliographic Entry**

|  |  |
| --- | --- |
| **Author:** | Bc. Hoa Vu Thu  Faculty of Science, Masaryk University  Department of ... |
| **Title of Thesis:** | Health data classification using neural networks |
| **Degree programme:** | Experimental Biology |
| **Field of Study:** | Computational Biology |
| **Supervisor:** | RNDr. Martin Komenda, Ph.D. |
| **Academic Year:** | 2019/2020 |
| **Number of Pages:** | ??+?? |
| **Keywords:** | Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword |

**Abstrakt**

V této bakalářské/diplomové/rigorózní práci se věnujeme ...

100 – 2000 znakov bez medzier

**Abstract**

In this thesis we study ...

|  |
| --- |
| **Místo tohoto listu vložte kopii oficiálního zadání práce bez podpisů.** |

**Poděkování**

Na tomto místě bych chtěl(-a) poděkovat ...

**Prohlášení**

Prohlašuji, že jsem svoji bakalářskou/diplomovou/rigorózní práci vypracoval(‑a) samostatně s využitím informačních zdrojů, které jsou v práci citovány.

|  |  |
| --- | --- |
| Brno xx. měsíce 2020 | ………………………………  Hoa Vu Thu |

**Content**

[1. Introduction 9](#_Toc25437973)

[Czech national health system and web presentation 9](#_Toc25437974)

[Natural Language Processing 9](#_Toc25437975)

[Explanatory/Research questions 11](#_Toc25437976)

[2. Methods 12](#_Toc25437977)

[CRIPS-DM 12](#_Toc25437978)

[Technologies? 12](#_Toc25437979)

[Software 12](#_Toc25437980)

[Libraries? 12](#_Toc25437981)

[NLP approaches and algorithms 12](#_Toc25437982)

[Neural networks 12](#_Toc25437983)

[Early years of Neural Network 12](#_Toc25437984)

[Biological background 13](#_Toc25437985)

[The main concept of NN 14](#_Toc25437986)

[Models for word representation 17](#_Toc25437987)

[Convolutional neural network ? 17](#_Toc25437988)

[Recursive neural network ? 17](#_Toc25437989)

[Reccurent neural network 17](#_Toc25437990)

[Models for Distributed representations of words 18](#_Toc25437991)

[Word2Vec 18](#_Toc25437992)

[Doc2Vec 20](#_Toc25437993)

[GloVe 22](#_Toc25437994)

[FastText 22](#_Toc25437995)

[Comparison of models 23](#_Toc25437996)

[Siamese architecture 24](#_Toc25437997)

[3. Result 26](#_Toc25437998)

[Data 26](#_Toc25437999)

[Training corpus 26](#_Toc25438000)

[Testing corpus 26](#_Toc25438001)

[Validating corpus 26](#_Toc25438002)

[Data preprocessing 26](#_Toc25438003)

[Modeling 26](#_Toc25438004)

[Evaluation 26](#_Toc25438005)

[Comparison of models 26](#_Toc25438006)

[Deployment 26](#_Toc25438007)

[Evaluation 26](#_Toc25438008)

[4. Discussion 27](#_Toc25438009)

[5. Conclusion 28](#_Toc25438010)

[References 29](#_Toc25438011)

# Introduction

## Czech national health system and web presentation

The Institute of Health Information and Statistics of the Czech Republic (IHIS) is an organisational component of the Czech Republic. It belongs to the Ministry of Health who delegated it to administrate the National Health Information System (NHIS). According to the Health Services Act the IHIS was specifically entrusted with the administration of:

* Data from the Program of statistical investigations of the Ministry of Health collected according to the Act on state statistical service
* National health registers including:
* the Czech National Cancer Registry
* the National register of hospitalisations
* the National register of reproduction health
* the National register of cardiovascular surgery and intervention
* the National register of joint replacement
* the National register of occupational diseases
* the National register of drug addict therapy
* the National register of injuries
* the National register of persons permanently excluded from blood donations
* the National register of autopsy and toxicological examination performed at forensic medicine departments
* the National register of health care providers
* the National register of health care professionals
* Data taken over from information systems on infectious diseases maintained according to Act on public health protection

All this data is used for monitoring health status of the population, the activity of health care providers, their economy and for obtaining the information about extend and quality of provided health services and for creation of health policy. NHIS is also designated for conducting and processing surveys on health status of population, on determinants, on the need and the consumption of health services, on its satisfaction and expenditure, for the needs of science and research in the field of the health.

The IHIS has currently several projects for example “Early detection of thyroid disease” in, or “Codification for Rare Disease”. For purpose of this thesis there will be introduced 2 of many projects and that is:

“**Methodological optimization and streamlining of the system of reimbursement of hospital care in the Czech republic”:**

This project was brought to the existence for the purpose of creating an information system of inpatient care reimbursement called CZ-DRG (diagnosis related group) and by its implementation there will be higher effectivity in management, controlling and optimization of distribution of financial flows in the health care.  
 The main goal is to build a long-term data, information and personnel base for the optimization and continuous cultivation of the system in the Czech Republic and to increase the predictive ability and effectiveness of reimbursement mechanisms for health care.  
 For achieving the main goal a Competence Centre and a representative network of reference hospitals had to be built and now it consists 49 hospitals, methodologies for evaluation of hospital cases, electronic and database tools for the implementation of the system had to be developed and verified in a representative network of reference hospitals, a new classification of hospitalization procedures had to be formed, the DRG classification system has to be redefined and a new coder’s manual had to be designed. The redefined DRG classification system can be found on <https://drg.uzis.cz/klasifikace-pripadu/web/>.

“**Centre for the Development of the Technology Platform of National Health Information System Registers, Modernization of their Content Extraction and Extension of their Information Capacity”**:   
 The target of this project is methodological development of NHIS, increase its information capacity and value and also development of necessary supporting software applications and components. It will also contribute to more efficient collection of data, better usage of information by the state administration and overall development of the system of departmental registries NHIS.

## Natural Language Processing

**Natural Language Processing (NLP) is an area of research and application that investigates how computers understand and process natural language text or speech. It is an interdisciplinary field of science that is consisted of computer science, linguistic, artificial intelligence and information engineering. It can be used for multiple application such as machine translation, named entity recognition or speech recognition.**

**The techniques that are used in syntactical analysis are:**

**Lemmatization**

**Morphological segmentation**

**Word segmentation**

**Part-of-speech tagging**

**Parsing**

**Sentence breaking**

**Stemming**

**Semantics refers to the meaning that is expressed by a text. It involves applying computer algorithms to understand the meaning and interpretation of words and how sentences are structured.**

**The techniques used id semantic analysis are:**

**Names entity recognition (NER)**

**Word sense disambiguation**

**Natural language generation**

* Information Retrieval([Google](https://www.google.com/) finds relevant and similar results).

Natural language processing (NLP) is an area of computer science and [artificial intelligence](https://www.indiumsoftware.com/machine-learning/) that deals with the interaction between computer and humans in natural language. The main purpose is to enable the systems to understand various language as humans do. It is the driving force behind [NLP products/techniques](https://www.indiumsoftware.com/text-analytics/) like virtual assistants, speech recognition, machine translation, sentiment analysis, automatic text summarization, and much more.

Natural Language Processing for Prolog Programmers - Michael A. Covington Artificial Intelligence Programs he University of Georgia Athens, Georgia – BOOK

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on a set of theories and a set of technologies.  
 Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. (<https://surface.syr.edu/cgi/viewcontent.cgi?referer=https://scholar.google.cz/&httpsredir=1&article=1019&context=cnlp>)

History

The 1950 was the beginning of the NLP as the intersection of artificial intelligence and linguistics but without information retrieval (IR – introduction to IR in the article Manning C., Raghavan P Schuetze H. Introduction to Information Retrieval . Cambridge, UK : Cambridge University Press, 2008). The first simple approaches were defeated by homographs which are words identically spelled with multiple meanings.   
(Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, Natural language processing: an introduction, Journal of the American Medical Informatics Association, Volume 18, Issue 5, September 2011, Pages 544–551, <https://doi.org/10.1136/amiajnl-2011-000464>)

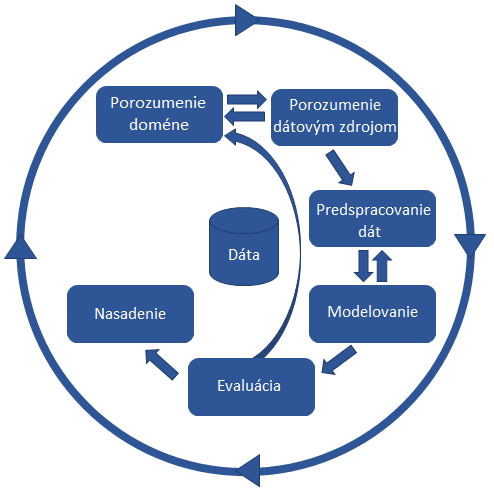
https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

## Explanatory/Research questions

# Methods

## CRIPS-DM

This thesis was processed by a data mining method Cross Industry Standard process for Data Mining (CRISP-DM). This method is consisted by 6 steps. These steps are connected mutually.



## NLP approaches and algorithms

### Neural networks

In this chapter it going to be introduces neural networks and it is going to be demonstrated how it works on the simplest model of neural network – perceptron. The information from this chapter derive mainly from (Neural Networks: A Systematic Introduction : Raul Rojas  
<http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=3ED2652AA9B56E647E98C9A4D44EA4F6?doi=10.1.1.18.493&rep=rep1&type=pdf>  
)

Neural network (NN), also called artificial neural network (ANN) is a subset of machine learning algorithms. They are largely used in image-recognition. They can reach accuracy up to 90% and no other machine learning algorithms has better performance(citacia). In recent years they became quite popular also in natural language processing (citacia).

### Early years of Neural Network

The first neural network was presented by Warren McCulloch and Walter Pitts in 1943. Those two men were inspired by the biological neural cells – neurons and the way how it works and they created a simple model of NN . (A logical calculus of the ideas immanent in nervous activity – W. S. McCulloch, W.H. Pitts). In 1969 was published a book called Perceptrons by Minsky and Papert, where was shown the deficiencies of perceptron model – which is the siplest model of NN (example). The research in neural networks was not funded anymore and only few researchers continued in their works. After some theoretical result, the interest in NN was renewed. ( napist ako sa to prestalo pouzivat a potom preco sa tozase zacalo pouzivat - http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=3ED2652AA9B56E647E98C9A4D44EA4F6?doi=10.1.1.18.493&rep=rep1&type=pdf)

### Biological background

The nervous system is a net of neurons connected by synapses.

Each neuron consists of:

a body cell (also called soma) – it contains a nucleus

an axon – delivers signals to another neuron

dendrites – receive incoming signals

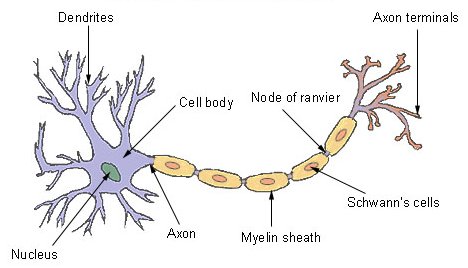
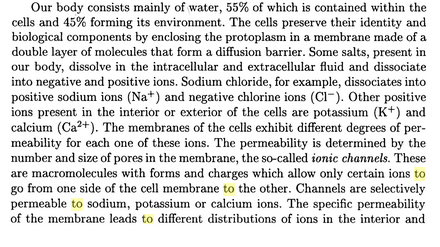
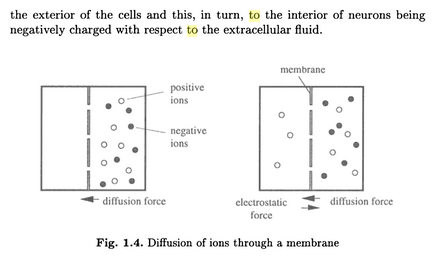


Figure: A simple structure of neuron (Source: [Quasar Jarosz CC BY SA 3.0, via Wikimedia Commons](http://en.wikipedia.org/wiki/File:Neuron_Hand-tuned.svg))

The three main parts of neuron correspond to a single unit of neural network.

Human body contains about 55% of water. Neuron transmit information via electrical signals.



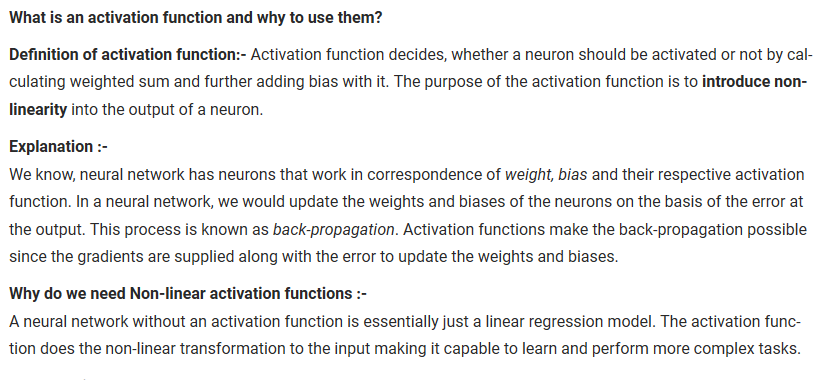


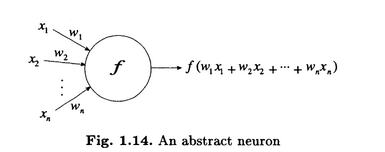
..... about action potential – physiologz of neuron

### The main concept of NN

As it is mentioned above the NN is composed of units – neurons. Several units create a layer. There are 3 types of layers: input, output and hidden layer, which can have multiple layers. Between each layer are sets of weights. In general, the input units receive a signal from the outside of the NN, they process this signal and sent it to the follow-up units – the hidden units, which process the received signals in the same way as the input unit. This process is repeated within the hidden layer until the signals is received to the output units. It is assumed, that each unit provides additive contribution to the input of the unit which is connected. The input to the unit is linear combination of the weight and the output of the connected unit from previous layer which is connected to plus bias. …. About activation function decides whether a neuron will be activated or not. The purpose of the activation function is to introduce nonlinearity into the output of a neuron.

Activation function is a primitive function which transforms input of the unit into a defined output. ….





Blablabla

Three elements are important to define a neural network:

* Structure of the nodes
* The topology of the network
* The learning algorithm used to find the weights of the network

In the training process, every iteration consists of several steps:

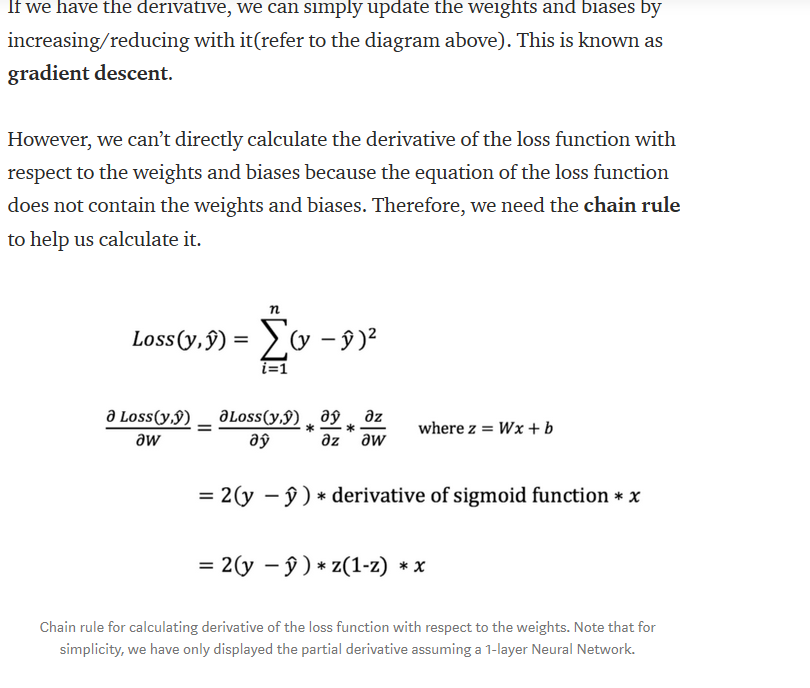
* Feedforward – prediction of the output
* Backpropagation – updating the weights and biases

**Loss function** is used to evaluate the goodness of the prediction from the trained neural network. There are many loss functions that can be used, for example the sum-of-squared error defined as:



Where y is … y is …. N is number of training examples. This is simply a sum of the differences between predicted value and the actual value.

The key in training neural network is to find the best set of weights and biases that minimizes the loss function but it is also important to avoid overtraining the network.

**Backpropagation -** blabalbal  
gradient descent – updating weights and biases by increasing or reducing with its derivative? 

[https://doi.org/10.1016/B978-0-12-741252-8.50010-8](https://doi.org/10.1016/B978-0-12-741252-8.50010-8" \o "Persistent link using digital object identifier" \t "_blank) - clanok, najst si h o doma

### Models for word representation

Distributed word representation = word vectors

<http://ruder.io/a-review-of-the-recent-history-of-nlp/>

distributional hypothesis – the idea that the meaning of a word is captured by the contexts on which it appears. Thus, the quality of the word vectors directly depends on the amount and quality of word vectors directly depends on the amount and quality of data they were trained on. (http://www.lrec-conf.org/proceedings/lrec2018/pdf/627.pdf)

### Reccurent neural network

Sequential architecture

Sequence modeling like language modeling as it requires flexible modeling od context dependencies

**Long short term memory (LSTM)**

**Gated recurrent unit (GRU)**

### Models for Distributed representations of words

Nowadays distributed word representations, which will be explained in this chapter, have been widely used in NLP. (http://www.lrec-conf.org/proceedings/lrec2018/pdf/627.pdf)

Sentence vector = paragraph vector = document vector – arbitrary-length abstract representation of the contextual meaning of a particular document type.

<https://arxiv.org/pdf/1405.4053.pdf>

Training, for successful application of distributed word representations it has to have a large corpora for training and using these pretrained models in downstream tasks

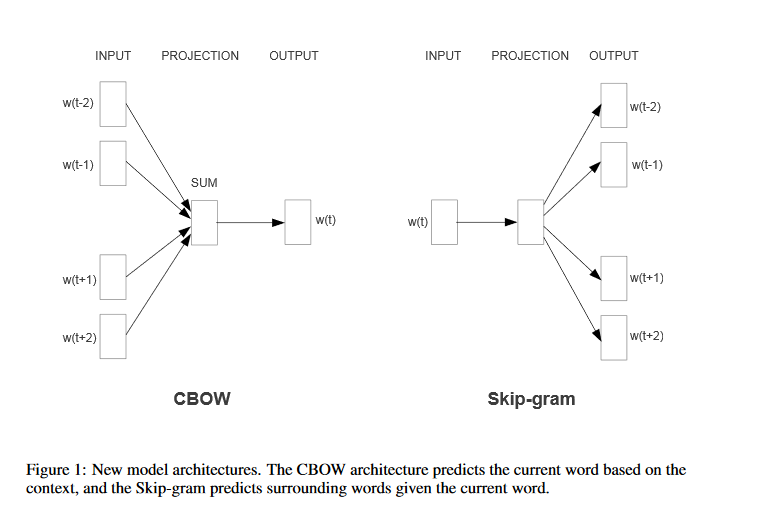
### Word2Vec

Multiple sets of features for automatic genre classification of web documents no NN  
https://www.sciencedirect.com/science/article/abs/pii/S0306457304000676

Word2Vec are models used in natural language processing (NLP) to produce a word embedding. In general, word2vec is a two-layer neural network that is used to recreate linguistic context of words. As its name refer, word2vec transforms a corpus of text into a vector space with a big dimension and every word is assigned to a vector in the space. These vectors are called word vectors. Two words vector are close to each other when they share a similar context or meaning. Basically, it treats each words in corpus like an atomic entity and generate a vector for each word.

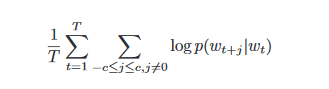
Word2Vec use two architecture: **continuous bag-of-word (CBOW) and continuous skip-gram**.

Continuous bag-of-words predicts the current word from the context. The order of context words does not affect prediction. In contrast, the continuous skip-gram uses the current word to predict the surrounding words and the weighs of words near to the current word are greater than the weights more distant from the current word. Continuous skip-gram is slower that CBOW but it has a better performance in case of infrequent words.

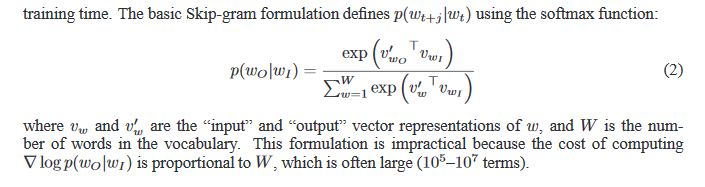


(<https://arxiv.org/pdf/1301.3781.pdf>)

**Skip-gram model** – this model search for word representations that help to predict the surrounding words in a sentence or document. Assume that we have a given sequence of training words *w1, w2, w3 … wT*. The skip-gram model maximizes the average log probability



Where *c* is the size of the training context (which can be a function of the center word *wt*). With the larger *c,* the higher accuracy can be obtained but the training time can be also increased.



(<https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>)

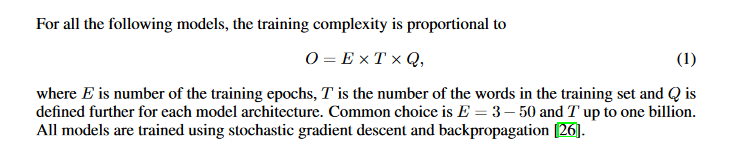
**Softmax**

**Hierarchical softmax**

**Negative sampling**

**Subsampling of frequent words**

The reason why is word2vec successful in word embeddings is poorly understood. The reasons for successful [word embedding](https://en.wikipedia.org/wiki/Word_embedding) learning in the word2vec framework are poorly understood. Goldberg and Levy point out that the word2vec objective function causes words that occur in similar contexts to have similar embeddings (as measured by [cosine similarity](https://en.wikipedia.org/wiki/Cosine_similarity)) and note that this is in line with J. R. Firth's [distributional hypothesis](https://en.wikipedia.org/wiki/Distributional_semantics). However, they note that this explanation is "very hand-wavy" and argue that a more formal explanation would be preferable.[[3]](https://en.wikipedia.org/wiki/Word2vec#cite_note-explain-3) Levy et al. (2015)[[16]](https://en.wikipedia.org/wiki/Word2vec#cite_note-16) show that much of the superior performance of word2vec or similar embeddings in downstream tasks is not a result of the models per se, but of the choice of specific hyperparameters. Transferring these hyperparameters to more 'traditional' approaches yields similar performances in downstream tasks. Arora et al (2016)[[17]](https://en.wikipedia.org/wiki/Word2vec#cite_note-17) explain word2vec and related algorithms as performing inference for a simple generative model for text, which involves a random walk generation process based upon loglinear topic model. They use this to explain some properties of word embeddings, including their use to solve analogies.  
The word embedding approach is able to capture multiple different degrees of similarity between words. Mikolov et al. (2013)[[18]](https://en.wikipedia.org/wiki/Word2vec#cite_note-18) found that semantic and syntactic patterns can be reproduced using vector arithmetic. Patterns such as “Man is to Woman as Brother is to Sister” can be generated through algebraic operations on the vector representations of these words such that the vector representation of “Brother” - ”Man” + ”Woman” produces a result which is closest to the vector representation of “Sister” in the model. Such relationships can be generated for a range of semantic relations (such as Country–Capital) as well as syntactic relations (e.g. present tense–past tense)



(https://arxiv.org/pdf/1301.3781.pdf)

### Doc2Vec

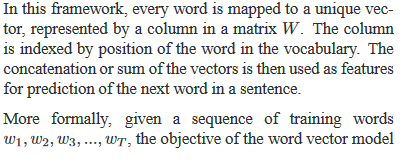
<https://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=159497>

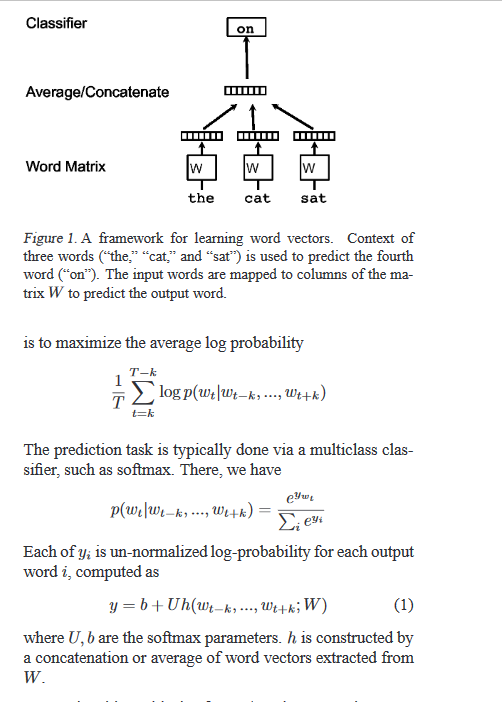
<https://towardsdatascience.com/multi-class-text-classification-with-doc2vec-logistic-regression-9da9947b43f4>

doi>[10.1145/1390156.1390177](https://doi.org/10.1145/1390156.1390177) A unified architecture for natural language processing: deep neural networks with multitask learning

Paragraph vector – unsupervised framework, it learns continuous distributed vector representation for pieces of texts (sentences or even documents). The vector representation is trained to predict words in a paragraph. The paragraph vector is concatenated with several word vectors from a paragraph and predict the following word in the given context. Words vectors and paragraph vectors, which are unique among paragraphs in contrast with word vectors, which are shared, are trained by stochastic gradient descent and backpropagation. At prediction time, the paragraph vectors are inferred by fixing the word vectors and training the new paragraph vector until convergence.

**Distributed vector representation of words**

the main goal is to predict a word given the other words in a context. 



http://ruder.io/word-embeddings-1/index.html

<https://medium.com/@dnberol/natural-language-processing-with-doc2vec-1d297059ac76>

### GloVe

Global vectors or Glove leverage global statistical information contained in a document. It is built on two main methods – global matrix factorization and local context windows. **Global matrix factorization** is used to reduce large term frequency matrices, which usually represent the occurrence or absence of words in a document. When it is applied to term frequency matrices it is called latent semantic analysis (LSA). Two methods are used in **local context windows:** CBOW and skip-gram that are mentioned in the chapter about word2vec.

### FastText

Recently, deep neural networks have become very popular for text processing. These models have very good performance but they can be slow to train and test which limits using very large datasets. FastText uses a hierarchical classifier which reduces the time complexities of training and testing (from linear to logarithmic with respect to the number of classes). It also exploits the fact that class are imbalanced by using Huffman algorithm. FastText was developed by Facebook team.

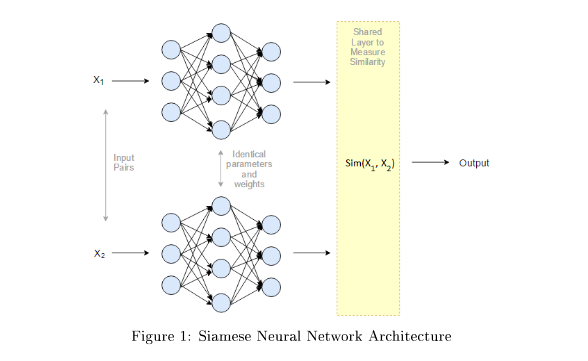
Essentially it is an extension of word2vec but instead of considering whole words it considers sub-words. It treats each word as composed character of n-grams. The word vector is made of a sum of this character n grams. It generates better word embeddings for rare words. It can construct the vector for a word from its character n grams even if word does not appear in training corpus

### Comparison of models

**Neural networks for text comparison**

### Siamese architecture

<http://ceur-ws.org/Vol-2028/paper8.pdf>



The objective of the siamese architecture is not to classify input images, but to differentiate between them.

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2789-2>

cosine of ht enagle between two feature vector is calculated and this represents the distance value

<https://arxiv.org/pdf/1904.11968.pdf>

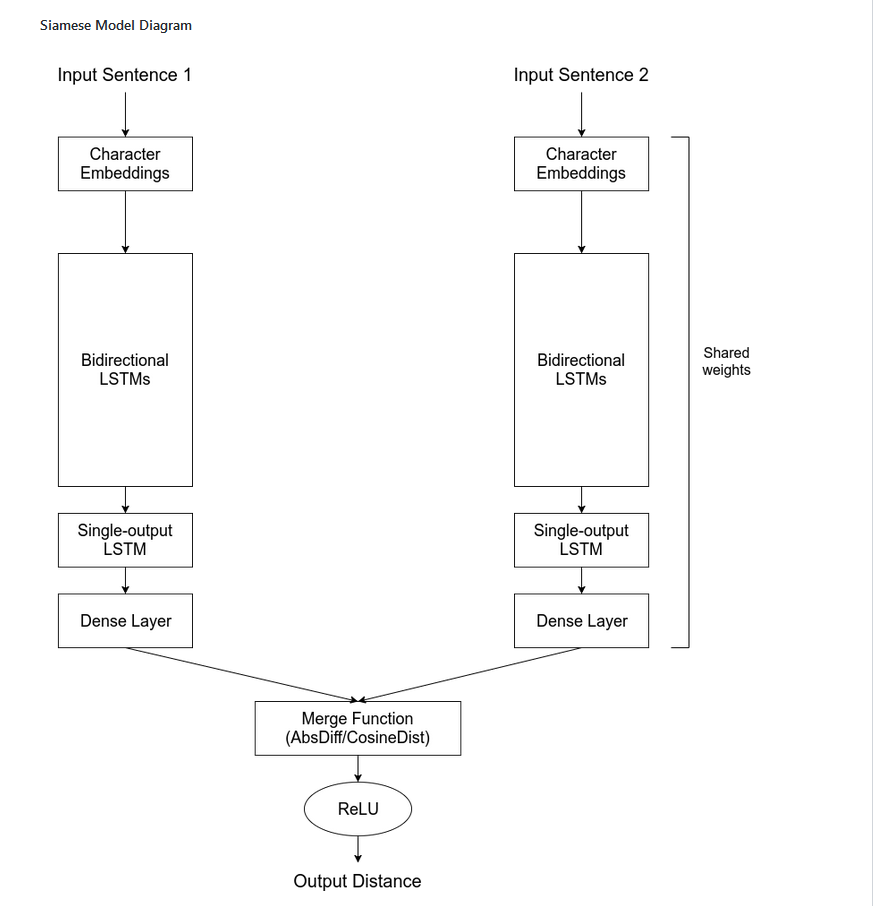
## Technologies?

## Software

Python 3.

## Libraries?

?



<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

<http://colah.github.io/posts/2015-08-Understanding-LSTMs/?source=post_page-----37e2f46f1714---------------------->

<https://github.com/amansrivastava17/lstm-siamese-text-similarity>

doc2vec

<https://praveenbezawada.com/2018/01/25/document-similarity-using-gensim-dec2vec/>

# Result

steps of CRIPS-DM

## Data

The corpus used for training models in this thesis was from (IDK YET) and for testing the model was from <https://drg.uzis.cz/klasifikace-pripadu/web/> and <https://reporting.uzis.cz/>. Those three websites are created by Institute of Health Information and Statistics of the Czech Republic. There is information about the health care in Czech republic in Czech language.

### Training corpus

### Testing corpus

Crawled data from <https://drg.uzis.cz/klasifikace-pripadu/web/> and <https://reporting.uzis.cz/>. The data were extracted using Xpath.

### Validating corpus

## Data preprocessing

## Modeling

## Evaluation

### Comparison of models

## Deployment

## Evaluation

# Discussion

# Conclusion

# References

# Online references

URL 1: <https://www.uzis.cz/> The Institute of Health Information and Statistics of the Czech republic [online] [cit. 23.November 2019]